DOCUMENT RESUME

ED 281 878 TM 870 282

AUTHOR Reynolds, Arthur J.; Oberman, Geraldine L.
TITLE An Analysis of a PSAT Preparation Program for Urban Gifted Students.
PUB DATE Apr 87
NOTE 41p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academically Gifted; Classroom Observation Techniques; *College Entrance Examinations; Comparative Testing; Grade 10; High Schools; Mathematics Tests; Pretests Posttests; *Program Evaluation; Questionnaires; *Test Coaching; Test Results; *Test Wiseness; Validity; Verbal Tests
IDENTIFIERS *Preliminary Scholastic Aptitude Test; Scholastic Aptitude Test; *Test Preparation Program Gifted Talented Sophomore; Tests of Achievement and Proficiency

ABSTRACT

The history of test coaching has shown that positive benefits can be produced from short-term orientation, practice, and drill. Coaching on the Scholastic Aptitude Test (SAT) has been especially scrutinized for efficacy and has shown modest program effects, despite the limited scope and duration of such programs. The present investigation expands the focus of coaching programs by assessing the effects of an intensive 63-hour Preliminary SAT (PSAT) verbal and math program on identified urban gifted students conducted in the summer of 1986. Based on a pre- posttest comparison group design, results showed a significantly positive math program effect for two of three sites tested and an overall math coaching effect of 4.2 points. Analyses of covariance results confirmed the significance of this math effect. However, a verbal coaching effect was not found as comparison group students outperformed summer program students from pre- to posttest. The observed math coaching effect was attributed to the program and was of greater magnitude than previous SAT/PSAT coaching studies. Given the strict adherence to practice, drill, and feedback on past PSAT's or SAT's, moderate program length, and instructors who are knowledgeable of SAT-type tests, coaching programs can be very effective in maximizing student scores. However, the emphasis on such testwiseness strategies may not improve the criterion performance of coached students and may also divert test focus away from the development of cognitive skills. (Author/JAZ)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED281878

An Analysis of a PSAT Preparation Program for Urban Gifted Students

Arthur J. Reynolds and Geraldine L. Oberman

Chicago Public Schools

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

A. Reynolds

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the Annual Meeting of the American Educational Research Association, Washington, D. C., April, 1987

Requests for copies should be directed to Arthur Reynolds, Department of Research and Evaluation, 1819 W. Pershing Rd. (5NW), Chicago, IL 60609

TM 870 282

BEST COPY AVAILABLE

Abstract

The history of test coaching has shown that positive benefits can be produced from short-term orientation, practice, and drill. Coaching on the SAT (Scholastic Aptitude Test) has been especially scrutinized for efficacy and has shown modest program effects, despite the limited scope and duration of such programs. The present investigation expands the focus of coaching programs by assessing the effects of an intensive 63-hour PSAT verbal and math program on identified urban gifted students conducted in the summer of 1986. Based on a pre-posttest comparison group design, results showed a significantly positive math program effect for two of three sites tested and an overall math coaching effect of 4.2 points. Analyses of covariance results confirmed the significance of this math effect. However, a verbal coaching effect was not found as comparison group students outperformed summer program students from pre- to posttest. The observed math coaching effect was attributed to the program and was of greater magnitude than previous SAT/PSAT coaching studies. Given the strict adherance to practice, drill, and feedback on past PSAT's or SAT's, moderate program length, and instructors who are knowledgeable of SAT-type tests, coaching programs can be very effective in maximizing student scores. However, the emphasis on such testwiseness strategies may not improve the criterion performance of coached students and may also divert test focus away from the development of cognitive skills.

Introduction

This evaluation examines the fourth year (1986) of the Test Preparation Program for Gifted and Talented Sophomores (TPPGTS). The TPPGTS is a six-week, 63-hour intensive coaching program developed to teach high achieving students test principles and strategies necessary to do well on the Preliminary Scholastic Aptitude Test/National Merit Scholarship Qualifying Test (PSAT/NMSQT). The test is administered every year in October and is composed of verbal and quantitative items deemed relevant for college success. A central goal of the program is to improve the test performance capacity of enrolled students. The program ran from June 27, 1986 to July 31, 1986 and was located at three high school sites. Students were recruited citywide.

Research Perspective

Test coaching has had a short but controversial history in education. During the 1950's, early studies of the effects of practice and coaching on test performance were conducted in Great Britian and considered tests used to assign children to secondary schools (Anastasi, 1981). It was typically found that the degree of improvement depended on ability, educational experience, type of coaching, and the underlying characteristics of the test. Coaching programs for the Scholastic Aptitude Test (SAT) were also being initiated at this time. Early studies showed promising results despite pronouncements by the Educational Testing Service (ETS) and College Entrance Examination Board (CEEB) that such coaching was ineffective.

Since these early developments there has been extensive research on the effects of coaching on test performance, especially regarding the SAT. Dear (1953), of ETS, conducted one of the first SAT coaching studies and found that program students made significant mean gains over the control group of 12.9

points on the math subtest and 4.6 points on the verbal subtest. The 10-hour program was composed of drill and vocabulary development on items developed by ETS to be compatible with the SAT. Other studies of relatively short coaching programs conducted by ETS (Dear, 1958; French, 1955) showed even higher significant gain scores for SAT-coached students.

Pallone (1961) developed and analyzed a longer, more intensive SAT-verbal coached program (6 weeks, 90 minutes per day) that included training in vocabulary development, reading skills, and verbal analysis. Pre- and posttest results from two groups of 20 and 100 students showed gains of 98 and 109 points. Marron (1965) found verbal and math gain scores of 41.1 points and 67.6 points above expected scores for a seven-month, accelerated program for senior boys in private preparatory schools. However, both studies are limited by their lack of control groups.

A more recent study by the Federal Trade Commission (FTC, 1979) investigated the effects of commercial coaching programs on student test performance in two New York city coaching schools. Mixed results were found. A 40-hour program for junior and senior high students at one school showed significant verbal subtest improvement from the PSAT to the SAT of 30 and 27 points for first and second time test-takers, respectively, and 19 and 28 point gains on the math subtest. However, it was found that the coached group was significantly higher than the uncoached group in high school class rank, parent income, most recent English and math grades, and number of years of math taken.

Studies with randomized control groups have generally shown lower program effects than other programs, although the number of studies are few. One study, Roberts and Oppenheim (1966) analyzed a 7 1/2 hour PSAT coaching

program for urban and rural minority students. Results showed relatively modest gains over the control group from pre- to posttest on the verbal (1.44) and math subtest (.8 points). These gains correspond to 14 and 8 points, respectively, on the SAT. Other randomized studies have also shown significant results but of different magnitudes. Alderman and Powers (1980) found a verbal coaching effect of 8.4 SAT points at eight public and private high schools. Evans and Pike (1973) found a significant math coaching effect of 16.5 points for a 21-hour instructed and 21-hour homework-guided program.

To synthesize the plethora of SAT and PSAT coaching studies, two extensive literature reviews (Messick and Jungeblut, 1981; Slack and Porter, 1980) and a meta-analysis (DerSimonion and Laird, 1983) have been completed. Slack and Porter (1980) examined the results of 19 published SAT coaching studies, including those mentioned above, and found average gain scores of 29 and 33 points on the verbal and math subtests, respectively, although many of these studies did not include control groups.

Messick and Jungeblut (1981) confirmed the significant positive effects of SAT coaching studies but indicated that observed gains were reduced for controlled or randomized designs. They also found that program effects increased with student contact time. Using a logarithmic function analysis, they computed average score effects as related to the number of contact hours for 17 verbal and 14 math studies. From this model it was estimated that a verbal gain of 10 points would require 12 hours of instruction and a verbal gain of 20 points would require 57 hours. For the math subtest, 20 score points would require 19 hours of instruction, 30 score points would require 45 hours of instruction, and 40 points would require 107 hours. DerSimonian and Laird (1983), in a meta-analysis of published SAT test preparation programs,

found program effects differed by the type of evaluation model employed. Using many of the same studies as the above reviews, they found average program effects to be approximately 10 points on the verbal and math subtests for matched and randomized studies.

Thus, despite early pronouncements that special coaching can do little to improve SAT scores, research has consistently shown that statistically significant gains due to coaching can be expected from typical test preparation programs and gains generally increase with student contact hours. Based on the above studies, estimated average SAT coaching gains appear to be 10-20 points on the verbal and math subtest, although the verbal subtest is less amenable to coaching than the math subtest. However, it should be remembered that the studies cited varied considerably in the kind of instruction materials used and may not represent the best available means to producing maximum coaching gains. For example, the recent availability of past SAT forms provides more advantageous means for practice and drill on SAT items. The few randomized and/or highly controlled studies actually completed and their short duration further limits the above results. Most reported studies include graduating seniors in private high schools and are not generalizable to all student groups such as exceptional students, for example.

The present investigation expands the scope of past test preparation programs by evaluating the effects of a 63-hour intensive PSAT program on urban gifted students. Analysis of this program may help diversify the restricted range of programs and participants in determining coaching effects.

Evaluation Questions

The following evaluation questions were addressed in this study to measure program effectiveness:

A. Was the program implemented as designed? Were the instructional materials compatible with this design?

B. Did students participating in the summer program show greater improvement than comparison group students from pre- to posttest and was this improvement attributable to the program?

C. What was the students' assessment of the program and its materials? What did they learn from the program?

These questions will be of particular interest to groups who have a vested interest in this program including (1) students interested in preparing for standardized tests, (2) teachers devoted to test preparation instruction and facilitating testwiseness among their students, (3) counselors who are most interested in assisting students with academic and career plans, (4) school administrators in general who are crucial in communicating the relevance of test preparation to students, and (5) the educational and scientific community at large who can help facilitate and develop the evaluation of test preparation programs.

Evaluation Design

In order to test the central evaluation question, that of whether summer program students show greater growth on pre- and posttest PSAT tests than comparison group students, and untreated control group design was used (Cook & Campbell, 1979). This generally interpretable nonequivalent control group design tests criterion score differences between groups before and after the program. Determination of program effects is estimated to be the difference

between mean criterion score growth of experimental and comparison groups, that is, the mean score differences from pre- to posttest between summer program and comparison group students on PSAT verbal and mathematics subtests. While this evaluation design is generally interpretable it is less desireable than a control group design with controlled selection whereby students are randomly assigned to treatment conditions. Thus, selection differences between groups can bias results. The ethical question of whether to exclude students from participation in a program with documented positive effects precluded employment of this more powerful design.

Student Information

One hundred forty-eight (148) students enrolled in TPPGTS at three sites and were obtained from over 1400 students eligible to participate. Elgibility, as defined by program administrators, was restricted to students who scored at or between the seventh and ninth stanines in reading and math on the Fall 1985 Tests of Achievement and Proficiency (TAP), a standardized high school achievement test. Stanine scores are standard scores that range from one to nine with a mean of five and a standard deviation of two. Sixty-three (63) percent or 93 of the participating students completed the program and pre-posttests. One-hundred thirty-seven comparison group students were also obtained from the pool of eligible students through letters sent independently of the summer program. However, participation was restricted to eligible students who attended one of five schools selected to maximize the size and ethnic diversity of the group. Comparison group students, who may not have had time to participate in the program, were invited to take two practice

PSAT's before and after the summer program at Sites 1 and 3 in preparation for the October 1986 PSAT. Thirty-four (34) percent (47) of the students tested completed pre- and posttests.

TABLE 1

Summary of Student Background Information
(in percent of respondents)

Variable	Responses	Summer Program (N=93)		Comparison Group (N=47)
Sex	Male	49		44
	Female	51		56
Year in school (Sept. 1986)	Sophomore	21		18
	Junior	79		82
Home School	School A	30		65
	School B	10		26
	School C	9		7
	School D	11		0
Why enroll/ Why not enroll?	Interest	64	Vacation	54
	Do better on tests	76	School	26
	Nothing planned	32	Will study on own	26
	Parents	30	No letter	15
Plan to take October 1986 Test?	Yes	94		100
	No	6		0
Taken a PSAT before?	Yes	44		22
	No	56		78

Table 1 shows general background information for the matched summer program and comparison group students. As is shown, group distribution by sex, year in school, and intention to take the October 1986 PSAT were nearly equivalent. Approximately 80% of each group had just completed their sophomore year of high school. Both groups obtained a large proportion of

students from School A, although the summer program groups was more diverse in school representation. The major reasons cited by summer program students for enrolling in the program were wanting to do better on tests and general interest. The primary reason cited by the comparison group for not participating in the program was a conflicting summer schedule and not a lack of interest in the program. It should be noted that some students did not receive any notification about the program. In addition, summer program students appeared to have greater PSAT experience than their counterparts as nearly half the group has taken a PSAT before.

Table 2 displays fall 1985 TAP stanine scores by summer program and comparison group students who completed both pre- and posttests. As is shown, mean reading (7.9 and 8.0) and math (7.8 and 7.9) stanine scores are approximately equal between summer program and comparison group students, respectively. Thus, the academic achievement of both groups on these measures is equivalent. The distribution of TAP stanine scores between groups are also similar. On the reading subtest, comparison group students were evenly divided among stanine groups, while stanine eight summer program students had a slightly higher proportion of students (40%) than the comparison group. On the math subtest, differences between groups occurred at the seventh and eighth stanines as the highest proportion of summer program students were at the seventh stanine (40%), the reverse of the comparison group, where 43% were at the eighth stanine. It should be noted that TAP subtest and total test scaled scores were also compared between groups. Results were similar to the above stanine data as no significant differences existed on these measures.

TABLE 2

Distribution of TAP Stanine Scores
by Summer Program and Comparison Groups
(Summer Program N=93, Comparison group N=47)

Reading			Math		
Stanine	Summer Program	Comparison Group	Stanine	Summer Program	Comparison Group
7	33%	34%	7	40%	34%
8	40%	32%	8	36%	43%
9	27%	34%	9	24%	23%
Total	100%	100%	Total	100%	100%
Mean ($\bar{X}$)	7.9	8.0	Mean ($\bar{X}$)	7.8	7.9

Self-reported English and math grades were also obtained for group comparison and indicated that the proportion of students taking basic English and math courses were nearly equivalent between groups as no significant differences occurred. For example, 97% of summer program students have taken Algebra 1 as compared with 98% for comparison students. In addition, 92% of summer program students completed geometry which is similar to that of the comparison group (98%). Furthermore, self-reported mean English and math grades were similar as summer program students averaged 3.1 and 3.0 (out of a 4-point scale), respectively, as compared to 3.1 and 2.9 for comparison group students. Although the self-reported and unstandardized nature of these grade reports should not be forgotten, similar levels of achievement are suggested.

Teacher Selection

Nine teachers were selected for the TPPGTS; three at each of the three sites. Teachers at each site specialized in the content areas of language

arts, mathematics, or guidance. All teachers were selected based on recommendations forwarded to program administration, experience in test preparation, and general interest. Although teachers were experienced, only one had been involved in prior test preparation programs. Consequently, they were not extremely knowledgable about coaching the PSAT. One teacher aide was also provided at each site.

Program Description

The 75-hour summer program included approximately 63 hours of test preparation and was divided in three content areas: (1) language arts, (2) mathematics, and (3) guidance. It was organized around the content of the actual PSAT with special emphasis on test-taking strategies, testwiseness, and drill. Three main textbooks were used: (1) Barron's PSAT/NMQST Guide, (2) Strategies for Taking the SAT Test, and (3) Gruber's Inside Strategies for the SAT. At each site the program was divided into three groups of three classes on the basis of math PSAT pretest scores. Each site subgroup rotated between subject areas on a daily basis. Each class period was 50 minutes in length.

In language arts classes, students examined and reviewed vocabulary, sentence completion, verbal analogies, and reading comprehension sections of the PSAT. In the mathematics section of the course, basic mathematics, algebra, and geometry were primarily emphasized. The guidance area explored awareness and understanding of test-taking strategies, values clarification, and career choices. Future plans and goals were also discussed in this component.

With the exception of the primary textbooks, there was no uniform curriculum for each component. Concentrating on a particular area, each teacher was allowed the flexibility of selecting his/her own manner of

presentation and instruction. However, practice test-taking, drill, and testwiseness strategies were emphasized throughout.

Instruments

Student questionnaire. This 41-item self-report questionnaire assessed the content, instructional materials, and student understanding of testwiseness principles of the program. Administered at the end of the program, survey items were coded on a four-point Likert scale and were composed of items about general program assessment (i.e., "I learned test-taking strategies that I did not know before"), the effectiveness of materials ("How effective was Inside strategies for the SAT."), and understanding specific testwiseness principles ("In this program I learned that I should usually guess when I don't know the answer to a problem."). The higher an item was rated the more positive the particular dimension was assessed.

Teacher questionnaire. A 9-item multiple choice and open-ended questionnaire was completed by program teachers and asked them to document skills taught in the program, instructional strategies and evaluation criteria used, and problems encountered in implementation. Teachers also assessed the effectiveness of the instructional materials used in the program as well as its strengths and weaknesses. Recommendations for future test preparation programs were also solicited.

Classroom observations. To assess the implementation of the test preparation program, classroom observations were conducted at the three sites by two staff members from the Department of Research and Evaluation. Each subject area at the sites were visited once at the beginning of the program with the exception of one subject, which was observed twice during two different class periods. The subject areas of the double observations were

different at each site so the number of total observations within each subject area were equivalent. These descriptive observations documented various classroom variables including (1) the compatibility of the classroom activities and the program outline, (2) student time-on-task, (3) type and length of classroom activities, (4) materials used, and (5) the classroom learning climate. At the completion of each observation, brief interviews were completed with teachers that encompassed reactions to the materials, student reactions to the content, and the ability of students.

The PSAT. To determine student academic growth of program students past PSATs were used, specifically forms S and T of the Fall 1985 PSAT. These tests were administered at the beginning and end of the program to both summer program and comparison groups in alternate forms. It should be noted that Site 1 students were posttested with the October 1983 PSAT. The PSAT is a standardized achievement test given to high school sophomores in preparation for the SAT. According to the College Entrance Examination Board (CEEB) (1985), the sponsor and governing body of the PSAT and SAT, the PSAT is "a multiple-choice test that measures developed verbal and mathematical reasoning abilities important for academic success in college. It assesses ability to reason with facts and concepts rather than the ability to recall and recite them". Items assess reading comprehension, word meaning, sentences completion, basic arithmetic, algebra, and geometry. It is highly correlated with the SAT. There are 115 questions on the test, 65 in verbal and 50 in the mathematics section. Raw scores are corrected for guessing and converted to linearly derived scaled scores ranging from 20 to 80 with a mean of 50. Testing time is one hour and 40 minutes.

Procedure

Eligible students were sent letters in May 1986 inviting them to participate in the program. Although many eligible students did not receive these letters, those who enrolled in the 6-week program were assigned to one of three sites, depending on which site was closest to their home or regular school. Students at each site were divided into three groups on the basis of their pretest scores and subsequently rotated between three teachers, one in each of the three content areas of language arts, mathematics, and guidance. On June 16, 1986 second letters were sent to eligible students at five high schools inviting them to take two practice tests as comparison group students at Site 1 or 3. This was practical for many students who did not have time to enroll in the program. All students, summer program or comparison group, were administered alternate-form pre- and posttest PSATs before and after the program. Additional background information was also obtained. During the program, student and teacher questionnaires were completed by participants and classroom observations were conducted.

Results

Program Implementation

The implementation of the TPPGTS was assessed by classroom observation and ascertained the degree to which it was enacted. Based on nine observations conducted for the program, three at each site, it appeared the program was implemented as intended. Observational records indicated that the vast majority of students participated in classroom activities and were on-task during class.

Instructional activities were diverse and comprehensive. For example, during the 456 minutes of actual class time observed (approximately 51 minutes

per observation), 36% of the instructional activities were devoted to independent or small group completion of assignments, 32% to listening to lectures or demonstrations, 23% to classroom discussion and recitation in response to teacher questions, and 9% to other group-related activities such as correcting answers to test questions. In addition, completing assignments independently or in small groups was the most frequent instructional activity as 6 of the 9 observations noted this occurence.

The instructional activities observed consistently agreed with the outline and goals of the course. English classes primarily included review of vocabulary words and their root meanings, and test practice. A variety of supplemental materials were also used to satisfy this objective including a word study guide, How to Ace the SAT, and The College Prep Game.

In the math section, lecture, classroom discussion, and practice test-taking was primarily employed. Teachers commonly explained the rationale behind correctly answering geometry and algebra items. In addition to the primary texts, other materials included a test preparation workbook, Mastering the SAT and past PSAT math exams. For example, the math instructor at Site 1 employed test practice and drill techniques from old PSAT's to familiarize students with the PSAT.

Guidance component activities appeared to fall into two categories, (1) reviewing test-taking strategies and (2) exploring values, college, and career choices. While reviewing test-taking strategies was most often handled through lecture and class discussion, the latter activities were instituted as group and class discussion activities.

Classroom observers also assessed the learning climate of the program. Results showed that on a four-point scale, student behavior was adequately

controlled in the classroom (3.8), students were task-oriented throughout class (3.4), students participated in class (3.6), and teachers facilitated classroom learning (3.6). Classroom observers also noted that the diversity of the instructional format stimulated student involvement (3.7) and the primary instructional materials were not consistently used during class (2.3). This latter finding suggests the lack of a uniform instructional focus.

Thus, the TPPGTS appeared to be implemented as designed, as nearly all instructional activities and materials were used in accordance with the outline of the course. However, it should be noted that observations were few and conducted early in the program. To gage a comprehensive implementation of the program, additional observations should be conducted toward the end of the program.

Test Results

In order to assess the central question of the evaluation, that of whether summer program students made greater gains from pre- to posttest than the comparison group, a variety of statistical analyses were employed. To determine general within-group program effects, a paired-sample correlated t-test was used. This test is regarded as the most powerful for a paired, correlated sample (Hays, 1981). Second, to determine general PSAT program effects between summer program and comparison group students, Hotelling's (1931) T^2 statistic was computed across verbal and math subtests. This multivariate statistic is also widely regarded as the most powerful test for a p-variate, simultaneous comparison (Maracuillo and Levin, 1983). Third, to determine substantive program effects, univariate analysis of covariance was used for the verbal and math subtest. Although multivariate analysis of

covariance is generally more efficient, the verbal and mathematics subtests are considered separate and independent tests and will be analyzed along this line.

As shown in Table 3, paired t-test results indicated that total summer program students made significant mean pre- to posttest gains on the verbal ($t=2.52$, $p<.05$) and math ($t=7.82$, $p<.001$) subtests. Their math subtest gain score of 5.2 (48.8 to 54.0) was over 3 times greater than their verbal subtest gain score of 1.5 (42.8 to 44.3). Although the comparison group, suprisingly, had a higher pre-to posttest verbal gain them the summer program group (1.9 points) ($t=2.30$, $p<.05$), their math gain score of 0.6 was not statistically significant. Thus, while both groups gained similarily on the verbal test, the summer program group made substantial gains on the math subtest over and above the comparison group.

Table 3 also indicates that the pretest comparison group scores are higher than the summer program group for verbal and math subtest scores. However, independent t-test results indicated these average pretest scores were not statistically higher than summer program students, which further confirms the relative equivalence of groups before the onset of the program.

Table 3 further provides a PSAT score breakdown by program site. As shown, there were no significant verbal subtest gain scores at any site for summer program students. Gain scores ranged from 1.1 at Site 3 to 2.5 at Site 2. However, the comparison group at one of the two sites (Site 1), significantly improved their verbal score by the end of the program and had twice the gain of the summer program group. The reverse occurred at Site 3 as comparison group students had a lower verbal gain score (-0.1) than their summer program counterparts (1.1). There was no comparison group at Site 2.

Table 3

Mean Practice Pre- and Posttest Scores by Group and Site

Site/ Ocassion	VERBAL		MATH	
	Summer Program (St. Dev)	Comparison (St. Dev)	Summer Program (St. Dev)	Comparison (St. Dev)
SITE 1 (N=53/35)				
Pretest	42.0	43.9	48.7	50.4
	(8.9)	(6.1)	(9.0)	(7.6)
Posttest	43.3	46.5	56.1	50.7
	(8.5)	(7.7)	(8.2)	(7.2)
Gain	+1.3	+2.6**	+7.4***	+0.3
SITE 2 (n=19/-)				
Pretest	43.7	-	49.0	-
	(7.0)	-	(6.2)	-
Posttest	46.2	-	52.3	-
	(5.5)	-	(6.1)	-
Gain	+2.5	-	+3.3*	-
SITE 3 (n=21/12)				
Pretest	44.0	45.4	48.9	51.2
	(7.5)	(8.1)	(7.2)	(7.8)
Posttest	45.1	45.3	50.3	52.8
	(6.7)	(8.6)	(5.5)	(7.1)
Gain	+1.1	-0.1	+1.4	+1.6
TOTAL (N=93/47)				
Pretest	42.8	44.3	48.8	50.6
	(8.2)	(6.6	(8.1)	(7.6)
Posttest	44.3	46.2	54.0	51.2
	(7.6)	(7.8)	(7.6)	(7.2)
Gain	+1.5*	+1.9*	+5.2***	+0.6

* $p < .05$
** $p < .01$
*** $p < .001$

In regard to the math subtest, two of three sites showed significant pre-posttest gain scores. Site 1 students were primarily responsible for the total group gain as they improved 7.4 points from pre- to posttest ($p<.0001$). Site 2 and 3 students gained 3.3 and 1.4 points respectively. Math gains for comparison group students were minimal, but suprisingly, Site 3 comparison group students had a higher gain score than their summer program counterparts (1.6 to 1.4 points) It should be noted that individual site results should be regarded with caution, especially at Sites 2 and 3, because of relatively small sample sizes. Despite this caution, it appears the program had a differential impact across sites as both teachers and content were different. These differences should be kept in mind when interpreting program effects.

Given the overall statistically significant math gain scores of summer program students, especially at Site 1, between-group analyses were conducted to determine substantive program effects. Results indicated a significant overall multivariate difference between summer program and comparison group students ($T^2=15.70$, $p<.0001$). Not surprisingly, there were no significant differences between groups on the PSAT verbal subtest after accounting for verbal and math pretest scores. In fact, the comparison group had a larger gain score than the summer program group (1.9 to 1.5 points). However, significant differences between groups on the PSAT math subtest remained even after controlling for other background variables such as TAP reading, math, and total scores, number of math and English courses taken in high school, and overall high school math and English grades.

As a result of the above analyses, an estimation of program effects can be made on both the verbal and math PSAT subtests. This estimation, via analysis of covariance with multiple covariates, adjusts for preexisting

group differences on all included variables and generally increases the precision of estimates of program effects (Cook & Campbell, 1979). Independent subtest results indicated a positive treatment effect of 4.2 PSAT points on the math subtest after adjusting for group difference on the math practice pretest, 1985 TAP math scores, and 1985 TAP total scores. In other words, if the two groups started with the same math PSAT subtest, TAP math, and TAP total scores, the summer program group would have significantly outperformed the comparison group by 4.2 points on the PSAT math posttest. Unfortunately, this positive overall treatment effect cannot be generalized to the verbal subtest. In fact, results indicate a negative treatment effect on the PSAT verbal subtest. That is, after accounting for initial group differences on the math pretest TAP math subtest, and TAP total test the comparison group outperformed the summer program group by approximately 1 PSAT point, although this difference was not significant. Thus, verbal PSAT coaching had no apparent effect in raising participating student test scores.

It should be noted that many other variables were included in the analysis of covariance model in order to adjust for preexisting selection differences including the number of math and English courses taken and grades received, TAP subscales, and math and verbal pretest scores. These added variables did not improve the precision of treatment effect estimates or the prediction of verbal and math posttest scores.

PSAT Scores by Stanine Group

Analyses were also conducted to determine the extent that PSAT gain scores increased with stanine group. Although, as expected, math and verbal subtest scores increased significantly with stanine group (i.e., TAP reading stanine-verbal PSAT score correlation was .47 and TAP math stanine - math PSAT

score correlation was .75) gain scores did not always fit this pattern. For example, comparing verbal PSAT scores across TAP reading stanine groups, only ninth stanine students had significant pre-posttest gain scores (2.2 PSAT points). However, seventh stanine program students outgained eighth stanine student by two points (2.3 to 0.3). Regarding math subtest gain scores across comparable TAP math stanines, statistically significant gains were recorded for all three stanine groups, although seventh stanine students obtained the highest math gain score (6.7) compared to ninth (5.5) and eighth (3.3) stanine students. It should be noted seventh stanine students, scoring 13 to 15 points lower than ninth stanine students had a much greater opportunity to improve than eighth or ninth stanine groups.

Analysis within the Summer Program Group

Also of interest in this investigation were differences in PSAT performance by various student subgroups including gender, PSAT experience, and reasons for enrolling in the course. In regard to gender differences, male and female program students scored similarly on pre- and posttests as there were no statistically significant differences on verbal or math subtests. Male and female students had identical verbal pretest scores (42.8) and similar verbal posttest scores (43.7 and 44.8), respectively, while male students had slightly higher math pretest (50.0 to 47.4) and posttest scores (54.0 to 53.6). With the exception of the written expression TAP subtest, whereby female program students scored significantly higher than their male counterparts, groups were similar on all background variables such as year in school and TAP reading and math subtests.

PSAT experience appeared to have a greater effect on scores than gender type. Program students who had taken the PSAT before pretested significantly

higher on the verbal (t=2.60, p<.01) and math (t=4.19, p<.0001) subtests than student who had not taken the test previously. This gap was narrowed by the posttest as there were no significant differences between groups on the verbal subtest and smaller differences on the math subtest (t=2.19, p<.03). However, program students with no previous PSAT experience gained 7.1 points from pre- to posttest(46.1 to 53.2) compared to 3.8 points for experienced PSAT test-takers(53.0 to 56.8), although more experienced test-takers were older and more able than their PSAT counterparts.

It was also of interest to compare test scores by the nature of participation in the program. For example, it was hypothesized that students who enrolled in the program specifically to improve their PSAT scores would show greater gains than other students, such as those who entered the program only because their parents expected them to or thought it would be a good way to meet other students. Results indicated no significant differences in verbal or math subtest scores between students who enrolled only because their parents expected them to and those that enrolled otherwise. There were also no differences between students who enrolled because they wanted to do better on tests and those who enrolled because they didn't have anything else planned for the summer or thought it would be a good way to meet people.

Supplemental Reports of Program Effects

To supplement pre-posttest data, student questionnaires were completed during the last week of the program.

Student questionnaire. Questionnaire results of 101 students indicated that they rated the program quite positively on a scale from one to four, with four being the most positive. For example, of the 22 items assessing general

program content, students thought the English (3.4), math (3.4), and guidance (3.2) components were helpful in preparing for the PSAT, their vocabulary (3.4) and problem-solving (3.2) skills were improved, and as a result of the program they were better prepared to take the October 1986 PSAT (3.4).

Students agreed that many concepts were learned in the program including short-cuts to doing PSAT problems (3.4), using time wisely (3.5), the importance of knowing algebraic and geometric operations and understanding how the PSAT is designed (3.4). More specifically, students stated their level of agreement with a number principles taught in Inside Strategies for the SAT. Eighty-three (83) percent correctly disagreed or strongly disagreed "that the only way to do well on the vocabulary test is to memorize as many words as possible." Only 25% of the students strongly disagreed. In reference to the question regarding reading directions to the PSAT if already known, 70% of the students correctly disagreed or strongly disagreed that directions should be read in this instance. In response to a question about guessing, seventy percent of program students agreed or strongly agreed that they should usually guess when they don't know the answer to a problem. However, too many students disagreed (30%), suggesting they would not usually guess in this instance. Not so positively assessed features of the program was the relevance of the guidance component (2.8) newness of the materials (2.5), and meeting everyday (2.7). In addition, students indicated to a degree, that taking the practice tests was more effective than class instruction in preparing for the PSAT (2.7), although students were generally divided on the item.

Questionnaire responses were also compared between Site 1 students and those at the other sites in order to help explain gain score differences between sites. The only item found to be significantly higher in favor of

Site 1 students regarded guessing. Site 1 students learned to a greater degree than others that they should usually guess when they come to a item for which they didn't know the answer ($t=4.57$, $p<.0001$). Thus, the higher math gain scores of Site 1 students was not entirely explained through questionnaire responses.

Discussion

The major evaluation question addressed was whether summer program students outperformed comparison group students from pre- to posttest and if this performance can be attributed to the program. Results indicated a significantly positive program effect for the PSAT math subtest after accounting for math pretest scores, math TAP scores and total TAP scores. Site 1 results were the most dramatic as summer program students gained 7.1 PSAT points over their comparison group. Relatively small sample sizes precluded interpretation of program effects at the other sites.

In regard to the PSAT verbal subtest, there were no overall significant differences between summer program and comparison groups on the PSAT verbal posttest after accounting for differences on the verbal pretest, TAP reading and total test scores, high school English grades, number of high school English courses taken, and PSAT math pretest scores. In fact, results showed an overall negative program effect for the verbal subtest as comparison group students obtained higher gain scores from pre- to posttest than summer program students, especially at Site 1. The estimated overall program effect from analysis of covariance was -.7 PSAT verbal score points or a loss by the summer program group relative to the comparison group or approximately 7 points on the SAT scale. Although this loss is statistically nonsignificant,

it does indicate the effectiveness of the verbal coaching component of the program was poor.

The inconsistent results obtained across sites suggest that the program was implemented differently at each site. For example, the Site 1 instructor emphasized systematic practice-testing and feedback with past PSAT's that was not apparent at the other sites. Differences in teacher experience and knowledge of test preparation may have also played a role. The small sample sizes at Sites 2 and 3 may have further exacerbated observed results.

While these different results should be kept in mind, it should not dissuade us from interpreting verbal and math coaching effects. The primary question to be addressed here is the validity of attributing the positive and negative program effects to the test preparation program.

Establishing the Validity of Observed Program Effects

The nonequivalent control group design with pretest and posttest measures used in this study is a generally interpretable design for establishing the internal validity of observed results. Results are internally valid when program effects can be attributed directly to the program and alternative explanations of program effects are ruled out. The present design typically rules out many threats to internal and external validity that may bias estimates of program effects. Following Cook and Campbell's (1979) method of determining the internal and external validity of observed program effects, the primary validity threats are tested individually for plausibility so that the nature of test score gains can be found. Elimination of all or most threats to internal validity lends support to the attribution of test score gains to program effects while elimination of external validity threats would

support the representativeness of present findings to other student populations. The primary concern here will be the observed PSAT math coaching effects, especially at Site 1, since the primary program effect was observed here. However, interpretation of no program effects for the PSAT verbal coaching component will be discussed separately. The discussion will begin with internal validity threats.

Threats to Internal Validity Ruled Out

Based on the design and results of the study, the following threats as described by Cook and Campbell (1979) may be ruled out:

History. This threat occurs when criterion scores of an experimental group are influenced by forces outside the context of the program such as other people or other instruction. The present control group design generally accounts for this threat as it is assumed outside forces are influencing both treatment and comparison groups equally. This potential effect would then be controlled in the pre- posttest data. History is a special concern in education because students are continuously exposed to an amalgamation of instruction programs, all of which may influence each other. In the present study, this threat is further neutralized by the relative short duration and isolated conditions (summer school) of the program. History effects usually occur most frequently with longer running programs.

Maturation. This threat is of concern when an observed treatment effect may be due to the general maturing process (i.e., growing older, wiser, or more experienced) rather than the program. As with history, maturation effects usually take place over a long period of time and would rarely occur over a six-week period. Further, the present design generally controls for

maturation effects and it is unlikely differential effects were operating between the groups.

Testing. Testing effects may occur when observed program effects are the result of participants becoming more familar with a criterion test such as when the same test is used for all testing occasions. Again, the use of a pre-posttest control group design usually accounts for this potential rival hypothesis because both groups would take advantage of such a situation. In the present study, this threat is further mediated by the use of alternative test forms from pre- to posttest.

Instrumentation. This threat typically occurs when the measuring instrument is changed in some way between the pre- and posttest or when groups exhibit "floor" or "ceiling" effects. Both effects appear to be controlled in this study. The alternate-form PSATs used produce identical linear standard scores that are statistically equated. "Floor" or "ceiling" effects were not an issue because mean test scores hovered around the middle range of the PSAT scale.

Statistical regression. This threat generally refers to the upward movement of a pretested experimental group to their population mean at the posttest which is mistaken for a treatment effect rather than a statistical artifact. The threat is most ominous when the treatment group has a much lower pretest score than the comparison group and criterion measures are unreliable. The latter condition is not plausible since the PSAT is a nationally standardized test that has high test-retest reliability. The former concern is also minimized because (1) the summer program and comparison groups did not perform signficantly different on the math pretest to render regression a major threat, and (2) the pattern of results obtained on the math

test, in which the lower scoring summer program group overtook the comparison group, in a crossover fashion by the posttest. As Cook and Campbell explain, this outcome reduces the liklihood of a regression alternative explanation because it is not reasonable to expect the summer program group to regress above the posttest score of the comparison group.

Mortality. Mortality is a threat to observed program effects when a substantial number of students drop out of the program after the pretest and it is found that those students who drop out have different characteristics than students who stayed in the program. When attrition is high sample representativeness is compromised and estimated treatment effects become biased. Although evaluation data were reduced to include only those students who completed pre- and posttests, consequently eliminating this validity threat, substantial reduction of the program sample is problematic. Sixty-four (64) percent of those students pretested participated to some degree in the program and took the posttest, a fairly positive retention rate. Of the comparison group, only 38% completed both pre- and posttests, indicating the interest and motivation of this reduced group in improving their test scores. However, it was found that those students who dropped out of the program after the pretest received significantly lower verbal and math subtest scores than the final summer program group.

Other major threats to internal validity ruled out. Other plausible alternative hypotheses to observed PSAT math coaching effects ruled out, in this study include (1) diffusion of treatment, (2) compensatory equalization of treatments, (3) compensatory rivalry, and (4) resentful demoralization. These threats are used usually to explain minimal or no program effects and involve contamination between experimental and comparison groups. Thus, they

are not directly applicable to math coaching results. The most likely threat of the four would be resentful demoralization, whereby the comparison group reacts negatively to their no treatment status and delibrarately lowers its test performance. This would have the effect of inflating the estimated positive program effect. However, given the exclusively voluntary nature of the comparison group, it is unlikely these students resented their no treatment status.

Threats Not Ruled Out as Explanations for Program Effects

Selection. Selection is a threat and potential explanation for observed program effects when there are preexisting differences between experimental and comparison groups that cannot be measured or controlled. It is especially a problem when recruitment procedures of groups are different and assignment of participants to experimental and comparison groups is not controlled. Although the use of analysis of covariance in the present study controlled the effects of some measured characteristics between groups such as practice pretests and achievement test scores, it cannot control for unmeasured characteristics that may be different between groups (Cook and Campbell, 1979). These other unmeasured characteristics may have influenced the positive math coaching effects obtained. This problem has been of particular concern in coaching studies (Messick and Jungeblut, 1981).

Interest in improving PSAT scores and motivation to do well appear to be the primary threats as a result of selection. The fact that summer program students participated in an in-depth test preparation suggests they were more interested and motivated in improving their scores than the comparison group. The magnitude of this effect is unknown, but surely must be considered in interpreting the validity of the observed math coaching effect. Fortunately,

though, this motivation/interest effect is minimized by two findings. First, comparison group students were obtained on the basis of their interest in improving their own PSAT scores. The choice to take two practice tests was completely voluntary and indicated they were also interested in improving their test scores. However, many students did not have time to enroll in the program or did not receive a letter inviting them to participate. It was not because they were not interested. In addition, nearly all comparison group students indicated they would be taking the October 1986 test.

Second, the plausible explanation that summer program students were more interested in improving their test scores may be offset by the fact that the comparison group was a higher achieving group than the summer program group. They scored higher on all achievement test measures and pretests. This information should also be taken into account.

Thus, the probability of the observed math coaching effect being the result of selection differences is reduced by the explicit interest of improving their PSAT scores and the possibility they are a higher achieving group. The magnitude of this effect remains uncertain, and if this threat exists at all, it is most likely a modest one, but it cannot be ruled out.

Interactions with selection. These threats occur when selection cannot be ruled out and resulting differences may be combining with other internal validity threats such as maturation, testing, and history. These threats are difficult to estimate for nonequivalent groups because all selection differences cannot be measured or obtained. Selection-instrumentation can be elimination quickly as both groups scored at approximately equal intervals on the PSAT, thus, results could be interpreted similarly. Selection-maturation effects are minimized by the relatively short

duration of the program and the finding that withing group variances decreased from the pretest to the posttest for both groups suggesting that the selection-maturation threat was at best minimal. In the latter case it is assumed that if selection-maturation is operating, then differential growth between groups should be occuring withing groups as well, and within-group variances do not indicate this occurance (Cook and Campbell, 1979).

However, two other interaction threats cannot be dismissed so easily, selection-history and selection-testing. Selection-history, the most serious threat, occurs most typically in disseminated treatment programs where sites receive different kinds of instruction, and program effects are concentrated at one particular site. Such is the case with the math coaching component, as the primary treatment effect was concentrated at Site 1. One must ask if their was some specific event or local history that enabled students at this site to gain over 7 points. The other two sites, although they were composed of much smaller numbers of students showed gains of only 3.3 and 1.4 points, respectively. In addition, all three groups started with nearly identical math pretest scores. Taking out math PSAT scores at the Site 1 reduces the average gain score from 5.2 to 2.3 and the unadjusted program effect from 4.6 to .7 points.

Thus, the apparent site-specific math PSAT program effect indicates that specific content information at Site 1 may have been responsible for the major program effect and not the program in general. As previously noted, teacher characteristics may have also played a part in the observed test score differences between sites. Although the teachers recruited for the program were experienced in their particular subject areas, the Site 1 instructor may have been more adept in preparing students for the PSAT.

Selection-testing. An ironic feature about this evaluation is that the effectiveness of a test preparation program is determined by tests. Thus, the summer program students are exposed to more test practice and past PSAT tests than their comparison group counterparts. This increased exposure to tests per se, rather than the instruction of the program may have been responsible for observed program effects. Although it can be argued that test practice and completion of former PSAT's are an intimate part of program instruction, determination of which component is the most effective in improving test scores is a highly relevant issue, and cannot be resolved in this study. Self-reported student questionnaire data indicated that 30% of those survey indicated that the tests were more important than the instruction in learning about the SAT.

Generalizability of criterion scores. While positive coaching effects were found, the nature of these effects and their meaningfulness is uncertain. As Anastasi (1981) has indicated, it is not clear whether test score gains due to coaching also result in improved criterion score performance. The major purpose of test preparation programs, especially on ETS tests, is to learn the test's structure, how it is designed, testwiseness strategies, and effective response patterns and not substantive content training. Thus, it is uncertain whether a student who scored 52 on the PSAT after a test preparation program would have the same performance capacity as someone who scored 52 without test preparation. The relatively small sample size of the study should also be considered, especially at Sites 2 and 3. Generalizability of such results should be made with caution.

Threats to External Validity

The purpose of a nonequivalent control group design is to eliminate most threats to internal validity and establish a program effect. It is not

particularly well designed to produce externally valid results or results that would be similar across other students and settings. There are three major threats to external validity as described by Cook and Campbell (1979) including the (1) interaction of selection and treatment, (2) interaction of setting and treatment, and, (3) reactive arrangements. The latter two threats appear to be ruled out in this study. The interaction of setting and treatment or treatment effects varying with the setting can be generally ruled out on the grounds that educational test preparation programs are intended for a homogenous setting, the classroom, and would probably vary minimally across such environments. Also ruled out is reactive arrangements, the probability that participants in other educational settings will react differently to the program in ways that change the magnitude of program effects. This threat is greatest in laboratory studies where results are obtained in contrived and artificial settings. This is not the case with classroom research as testing and program instruction is a regular feature of education.

Unfortunately, the most relevant threat to be ruled out, selection-treatment, cannot be. This external validity threat limits the generalizability of results when program/treatment students are not representative of students at large. Obviously, program students are a selective student group in regard to their motivation to do well on the PSAT and as test-takers. Program students were in the upper third of the TAP achievement test distribution, thus, results cannot be generalized to students testing lower than this restricted range. However, results indicate that coaching effects would be even greater for lower ability students or novice test-takers.

The Status of the Program

On the basis of the above discussion it is possible to interpret the observed math coaching effect. Even though most threats to internal validity may be ruled out as influencing observed program effects such as history, maturation, testing, mortality, instrumentation, and to a degree, selection, a general program effect across sites did not exist. The primary positive effect of the program was observed at Site 1 in which gain scores can be attributed to systematic practice testing and analysis of past PSAT's and possibly the teacher. Thus, the positive program effect is site-specific but substantial. The fact that students coached at Site 1 gained 7.1 PSAT or 71 SAT points over their comparison group strongly indicates the positive effect of practice testing, drill, and feedback. Limited samples at the other two sites preclude interpretation of program effects.

The overall estimated math coaching effect of 4.2 PSAT points (42 SAT points or 71 points for Site 1 students) is much greater than the 10 to 15 point gains reported in the SAT review literature (DerSimonian & Laird, 1983; Slack and Porter, 1980), although negligible results were found at the other comparison group site. The present results do not conform to logarithmic model of Messick and Jungeblut (1983) as it was calculated that a math score gain of 40 points would require approximately 107 hours of instruction. Approximately 30 hours of math instruction was included in the program, indicating significantly greater time-effectiveness than other studies.

The above results indicate two major conclusions to the evaluation: (1) the desireability of using old PSAT's as instructional materials almost exclusively for practice, drill, and feedback, and (2) the need to establish clearer and more uniform guidelines for instruction. The use of past PSAT

forms serve the relevant function of familiarizing students with actual PSAT items rather than approximated PSAT items, thus give students a greater sense on the type of items they can expect on the test.

The status of the verbal coaching program is not so positive as a small, negative program effect was found. Thus, the verbal component may have been detrimental to summer program students. This unusual result is difficult to explain as internal validity threats are not generally applicable. One possible explanation may be selection. Enrollment in the program favored those students who had not made summer plans. Consequently, comparison group students may have been more academically involved. Consistently higher PSAT and TAP test scores support the greater academic experiences of this group. However, it should be noted that verbal PSAT is less amenable to coaching than math as many studies have confirmed this finding (Alderman and Powers, 1980; Dyer, 1953; French, 1955). If the goal of coaching is to improve test scores only, then it should correspond as closely as possible to the test content. The public availability of past PSAT and SAT forms enables this compatibility to be strong.

Comparability with Previous Years

The site-specific math program effect, then, does not undermine the observed positive coaching effect. These results are consistent with the 1984 and 1985 findings in which program students gained 4.4 and 3.5 PSAT points, respectively (Chicago Public Schools, 1986), although comparison groups were not employed. However, present verbal subtest results are not consistent with 1985 and 1984 findings as it was found that program students gained 3.2 and 7.1 PSAT points, respectively. These observed differences may be at least partially explained by student and teacher differences over the past three years.

Future Directions

TPPGTS has consistently shown significantly positive math coaching effects the past thre years on the order of approximately 36 SAT points. In regard to verbal coaching, TPPGTS has produced mixed results. While the 1984 program resulted in a pre- posttest gain of 71 SAT points, present results indicate a small negative program effect. However, that present summer program students had a verbal gain of only 15 SAT points suggests a differential program impact from year-to-year. The fact that 1986 program students pretested four points higher than 1984 students partially accounts for these different results. Overall results, though, suggest that substantial SAT gain scores due to coaching are possible for programs of this length, especially in math.

Although reviews on SAT coaching have born out that significant gains due to coaching can be expected, programs have not been enthusiastically recommended (DerSimonian and Laird, 1983). The present results suggest to the contrary, especially regarding math coaching, and indicate that substantial gain scores can be obtained by above average students. It remains to be seen whether these results are generalizable to other students, although more positive gain scores would be expected for lower ability students(i.e., less than stanine 7).

However, given adequate instructor training and knowledge of the SAT, standardized instruction, and the most up-to-date instructional materials such as old PSAT's or SAT's, coaching gains can be further maximized. In addition, most reviewed coaching programs have been of relatively short duration (less than 30 hours), which may serve to underestimate programs of longer duration. It is imperative that studies be undertaken with more diverse populations and instructional approaches before the magnitude of coaching effects is confirmed.

Many issues, though, remain unresolved. As discused above, a test score influenced by coaching may not improve subsequent criterion behavior, thus, reduces the validity of a test (Anastasi, 1981). Future research is needed to resolve this issue more clearly. In addition, the necessity of employing rigorous evaluation methods in determining progrm effects cannot be overemphasized. Although coaching studies cannot be completely bias-free, the utilization of comparison group designs with controlled selection can minimize ambiguities about observed program efforts and provide more valid information about coaching. As this methodological problem has been noted previously for SAT coaching (Messick and Jungeblut, 1981), it is essential coaching studies employ such evaluation methods. Critical analysis of evaluation results is also mandatory, since alternative explanations to observed program effects must be addressed and interpreted.

Finally, the educational significance of improving college entrance examination test scores may be limited to the test score itself. SAT-type tests measure a very limited set of abilities, and if overemphasized, downplay the development of essential competency skills needed in society. Effective coaching also appears to concentrate more on testwiseness strategies and the idiosyncratic qualities of tests rather than substantive cognitive skills. Emphasis on instruction in problem-solving techniques and broad cognitive skills is clearly more relevant to student learning and adaption. Different from most coaching programs, training in broad cognitive is designed to improve criterion behavior rather than a test score itself. However, as long as the SAT is seen as the measuring stick of academic intelligence and college admissions, coaching will continue its wide appeal in education.

References

Alderman, D.L. and Powers, D.E. (1980). The effects of special preparation on SAT-verbal scores. American Educational Research Journal, 17, 239-253.

Anastasi, A. (1981). Coaching, test sophistication and developed abilities. American Psychologist, 36, 1086-1093.

College Entrance Examination Board (1985). A counselor's guide to helping students learn from the PSAT/NMSQT. Philadephia, Pa: Author.

Chicago Public Schools. (1986). A review of test preparation programs for gifted and talented sophomores, 1983-1985. Department of Research and Evaluation. Chicago, IL: Author.

Cook, T.D. and Campbell, D.T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Boston: Houghton Mifflin Company.

Dear, R.E. (1958). The effects of a program of intensive coaching on SAT scores (RB58-5). Princeton: Educational Testing Service.

DerSimonian, R. and Laird, N. (1983). Evaluating the effect of coaching on SAT scores: A meta-analysis. Harvard Educational Review, 53, 1-15.

Dyer, H.S. (1953). Does Coaching Help? College Board Review, 19, 331-335.

Evans, F.R. and Pike, L.W. (1973). The effects of instruction for three mathematics item formats. Journal of Educational Measurement, 10, 257-272.

French, J.W. (1955). An answer to test coaching. College Board Review, 27, 5-7.

Federal Trade Commission, Bureau of Consumer Protection, (1979). Effects of coaching on standardized admission examinations: Revised statistical analyses of data gathered by Boston Regional Office of Federal Trade Commission. Washington, D.C.: Author.

Hays, W.L. (1981). Statistics. New York: Holt, Rinehart and Winston.

Hotelling, H. (1931). The generalization of student's ratio. Annals of Mathematical Statistics, 2, 360-378.

Marascuilo, L.A. and Levin, J.R. (1983). Multivariate statistics in the social sciences. Monterey, Ca: Brooks/Cole.

Marron, J.E. (1965). Preparatory school test preparation: Special test preparation, its effect on College Board Scores and the relationship of affected scores to subsequent college performance. West Point, N.Y.: Research Division, Office of the Director of Admissions and Registrar, United States Military Academy.

Messick, S. and Jungeblut, A. (1981). Time and method in coaching for the SAT. Psychological Bulletin, 89, 191-216.

Pallone, N.J. (1961) Effects of short-term and long-term developmental reading courses upon S.A.T. verbal scores. Personnel and Guidance Journal, 39, 654-657.

Roberts, S.U. and Oppenheim, D.B. (1966). The effect of special instruction upon test performance of high school students in Tennessee (CB RdR66-7) No. 1 and ETS RB66-36). Princeton, N.J.: Educational Testing Service.

Slack, W.V. and Porter, D. (1980). The scholastic aptitude test: A critical appraisal. Harvard Educational Review, 50, 154-175.